REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Nov. 15, 1994	FINAL REPORT (6/1/91-9/30/94)

4. TITLE AND SUBTITLE
Bootstrap and Partitioning Methods

5. FUNDING NUMBERS
DAAL03-91-G-0111

6. AUTHOR(S)
Wei-Yin Loh

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Wisconsin-Madison
750 University Avenue
Madison, WI 53706

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 28679.15-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

ABSTRACT *(Maximum 200 words)*

The following problems are studied and their solutions found. (1) Bootstrap methods for confidence interval estimation of a binomial parameter and for model selection in linear regression. (2) Tree-structured algorithms for classification, piecewise-linear regression and generalized linear models, and proportional hazards regression for censored observations. (3) Asymptotic efficiency of tests following data transformations. (4) Identification of significant effects from unreplicated two-level factorial designed experiments. (5) Bounds on the asymptotic size of the likelihood ratio test of independence in a cross-classified table.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Bootstrap, recursive partitioning, decision trees, model selection	5
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Bootstrap and Partitioning Methods

Final Report

Wei-Yin Loh

November 15, 1994

U. S. Army Research Office

Grant Number DAAL03-91-G-0111

University of Wisconsin, Madison

1 Problems studied

1. Asymptotically consistent methods for variable selection in linear regression models.

2. Construction of second-order efficient bootstrap confidence intervals for a binomial proportion.

3. Tree-structured statistical methods for classification and nonlinear function estimation by recursive partitioning.

4. Asymptotic efficiency of the t-test following Box-Cox transformations and its effect on linear discriminant analysis.

5. Identification of significant contrasts in unreplicated two-level factorial experiments.

6. Bounds on the asymptotic size of the likelihood ratio test of independence in a two-way contingency table.

2 Summary of important results

The following results were obtained in each of the problems listed above. References refer to the list of publications in Section 3.

1. There are numerous methods for selecting the correct variables to use in a linear model. Well-known procedures include Mallows' C_p, Akaike's AIC, and cross-validation. It is shown in [9] that all of these methods are inconsistent in the sense that the probability of correct selection does not converge to unity as the sample size tends to infinity. A method that employs a heavier model complexity penalty is proposed and proved to be consistent.

   Bootstrap methods that attack the same problem are developed and analyzed in [8] and [10]. One bootstrap method improves upon an existing procedure by reducing its bias and variance. A second method selects a model based on bootstrap estimates of expected prediction error. This method is consistent even when the errors are heteroscedastic.

2. A method is developed for the construction of second-order efficient bootstrap confidence intervals for a binomial proportion. The method first smooths the data by convolution and then uses the PI's bootstrap calibration method (developed in 1987) to construct the interval. Second-order efficiency is proved via Edgeworth expansions. The results are reported in [11].

3. Algorithms for tree-structured classification [12], least-squares regression [5], generalized linear models [7], and proportional hazard regression [6] are developed. The algorithms employ recursive partitioning and cross-validation pruning. They are compared with existing tree-structured and non-tree-structured methods using real and simulated data sets. The results show that the new algorithms are as accurate as all the existing methods. They are fastest among tree-structured methods, with speed superiority up to several hundred times faster for moderate sample sizes. The speed advantage increases rapidly with increase in sample size and number of variables. Proofs of asymptotic consistency of some of the methods are obtained. The computer programs developed for this project are freely available from the PI.

4. The family of Box-Cox transformations is well-known to be a powerful data analytic tool. It is proved in [1] that the application of these transformations to the classical t-test yields a test with asymptotic relative efficiency bounded below by unity for all data distributions. Hence, at least in large samples, application of the Box-Cox transformations to the t-test is recommended. A similar, though less striking, result is obtained in the context of linear discriminant analysis in [3].

5. A standard but subjective technique for the analysis of data from two-level factorial experiments is Daniel's normal plot of the contrasts. It is shown in [2] that this technique is not invariant of the labeling of the factor levels. A method that does remain invariant and that identifies the significant contrasts in a completely objective way is proposed. Comparisons with other methods show that the new method has equivalent power. The method is now routinely taught in experimental design courses at the University of Wisconsin, Madison.

6. A long-standing problem in testing the independence of the rows and columns of data from a cross-classified table is which test statistic to use and with what minimum observed cell count. These questions are rooted in the convergence of the null distribution of the test statistic as the sample size increases. It is proved in [4] that in the case of the likelihood ratio statistic, this convergence can be highly non-uniform over the parameter space. This shows conclusively that it is futile to devise practical recommendations for the minimum cell count to use with the test.

3 List of publications

1. Bounds on AREs of tests following Box-Cox transformations (joint with H. Chen) (1992). *Annals of Statistics*, **20**, 1485–1500.

2. Identification of active contrasts in unreplicated factorial experiments (1992). *Computational Statistics and Data Analysis*, **14**, 135–148.

3. Application of Box-Cox transformations to discrimination for the two-class problem (with P. Qu) (1992). *Communications in Statistics*, **21**, 2757–2774.

4. Bounds on the size of the likelihood ratio test of independence in a contingency table (with X. Yu) (1993). *Journal of Multivariate Analysis.* **45**, 291–304. 1993.

5. Piecewise-polynomial regression trees (with P. Chaudhuri, M.-C. Huang and R. Yao) (1994). *Statistica Sinica.* **4**, 143–167.

6. Tree-structured proportional hazards regression modeling (with H. Ahn) (1994). *Biometrics.* **50**, 471–485.

7. Generalized regression trees (with P. Chaudhuri, W.-D. Lo and C.-C. Yang). *Statistica Sinica.* In press.

8. Bootstrap bias and variance reduction in the estimation of a model dimension (with X. Zheng). *Proceedings of the American Mathematical Society.* In press.

9. Consistent variable selection in linear models (with X. Zheng). *Journal of the American Statistical Association.* In press.

10. Consistent bootstrap variable selection in linear models (with X. Zheng). Submitted to *Annals of Statistics.*

11. Bootstrapping binomial confidence intervals (with X. Zheng). *Journal of Statistical Planning and Inference.* In press.

12. Split selection methods in classification trees (with Y.-S. Shih) (1994). Submitted to *Statistica Sinica.*

4 List of participating scientific personnel

The grant provided research assistantship support to the following five PhD students, with year of graduation given in parentheses.

1. Hongshik Ahn (Research assistant, PhD 1992)

2. Yu-Shan Shih (Research assistant, PhD 1993)

3. Ruji Yao (Research assistant, PhD 1994)

4. Xujie Yu (Research assistant, PhD 1994)

5. Xiaodong Zheng (Research assistant, PhD 1994)